FasterAnalytics for Biopharmaceuticals – A Case Study

Introduction

DecisionQ has developed FasterAnalytics, a unique analytics package that enables scientists and researchers to use sophisticated predictive analytics from their desktops. FasterAnalytics is fast and creates high quality, predictive models from data that enable day-to-day review of experimental data, real-time hypothesis testing, and rapid research prioritization decisions.

FasterAnalytics uses a modeling approach called Bayesian Networks to provide a mapping of the complex relationships in data, which can then be used to make high quality predictions. Users can:

- Get an instant global view of their data.
- Understand the driving factors in the data.
- Test hypotheses in real time in our model Explorer.
- Produce reports that can be exported to other applications.
- Make determinations that can help prioritize the use of scarce research resources.

Market Overview

The Biopharma industry spends billions of dollars annually on research. This market is served by an array of software vendors selling tools that enable researchers to bring better products to market faster. The pace of discovery requires companies to seek efficiencies in the research and candidate selection process in order to gain a competitive advantage and maximize the potential of candidates in the pipeline.

Value to the Customer

FasterAnalytics enables both experts and non-experts in statistics to discover and leverage knowledge from data at key 'bottle-neck' points in the research and discovery process. Examples include:

- Automated mapping of data where the targets are unknown to reveal a complex map of interaction pathways.
- Identification of both positively and negatively correlated relationships to a target variable.
- Ability to make predictions about the behavior of multiple factors and test hypotheses in real time.
- Dimensionality reduction of a large data set down to the most likely solution set.
- Prioritization of the most promising areas to test and fail-fast.

DecisionQ's predictive modeling software is designed for real-time environments. Bayesian models are highly effective at identifying experimental defects early on in the process and reducing the cost and time taken to reach a successful experiment.

Product and Technology

DecisionQ Corporation has produced a range of modules that perform data analysis, modeling, visualization, reporting, and decision optimization. FasterAnalytics modules include:

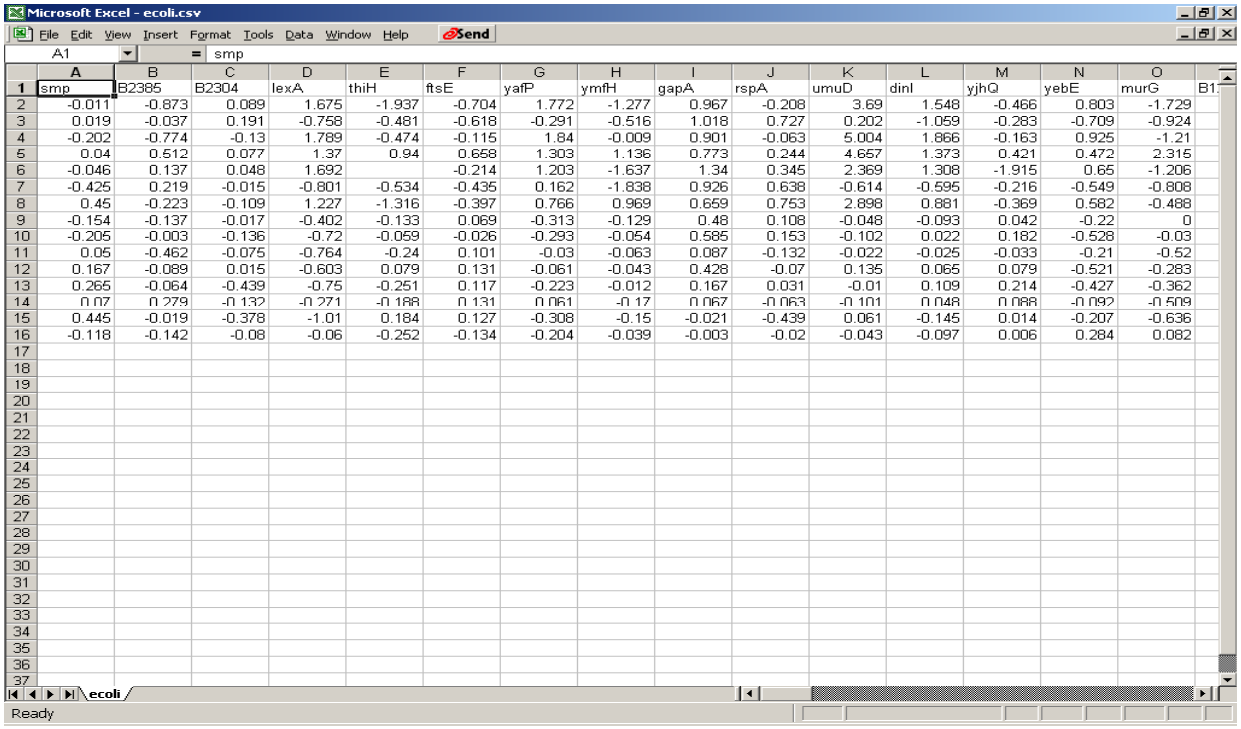
- *Discretizer.* Automatically configures the data for modeling.
- *Modeler.* Quickly creates a visual model of the data.
- *Explorer.* Allows real-time generation and testing of hypotheses.
- *Reporter.* Extracts insights and key points for inclusion in reports and presentations.

Using the System: A Microarray Example

The following is an example application of our software using a publicly available microarray data set. We have used a set with only 15 samples that has been used to study the SOS response in the Escherichia coli genome.[1] We have selected 122 genes, some of which are known to belong to the SOS response. FasterAnalytics prepared the data and built the model in this example in under 10 minutes.

To build predictive models, our learning engine requires the data to be in a flat tabular format. The data can be numerical, or variable character strings. Our software also handles missing values automatically and will either impute a value or treat missing values as a special category, at the user's discretion.

Figure 1: This example uses a micro array data set held in an Excel spreadsheet as shown below (Partial).


Microsoft Excel - ecoli.csv

File Edit View Insert Format Tools Data Window Help Send

A1 = smp

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	smp	B2385	B2304	lexA	thiH	ftsE	yafP	ymfH	gapA	rspA	umuD	dinI	yjhQ	yebE	murG	B1
2	-0.011	-0.873	0.089	1.675	-1.937	-0.704	1.772	-1.277	0.967	-0.208	3.69	1.548	-0.466	0.803	-1.729	
3	0.019	-0.037	0.191	-0.758	-0.481	-0.618	-0.291	-0.516	1.018	0.727	0.202	-1.059	-0.283	-0.709	-0.924	
4	-0.202	-0.774	-0.13	1.789	-0.474	-0.115	1.84	-0.009	0.901	-0.063	5.004	1.866	-0.163	0.925	-1.21	
5	0.04	0.512	0.077	1.37	0.94	0.658	1.303	1.136	0.773	0.244	4.657	1.373	0.421	0.472	2.315	
6	-0.046	0.137	0.048	1.692		-0.214	1.203	-1.637	1.34	0.345	2.369	1.308	-1.915	0.65	-1.206	
7	-0.425	0.219	-0.015	-0.801	-0.534	-0.435	0.162	-1.838	0.926	0.638	-0.614	-0.595	-0.216	-0.549	-0.808	
8	0.45	-0.223	-0.109	1.227	-1.316	-0.397	0.766	0.969	0.659	0.753	2.898	0.881	-0.369	0.582	-0.488	
9	-0.154	-0.137	-0.017	-0.402	-0.133	0.069	-0.313	-0.129	0.48	0.108	-0.048	-0.093	0.042	-0.22	0	
10	-0.205	-0.003	-0.136	-0.72	-0.059	-0.026	-0.293	-0.054	0.585	0.153	-0.102	0.022	0.182	-0.528	-0.03	
11	0.05	-0.462	-0.075	-0.764	-0.24	0.101	-0.03	-0.063	0.087	-0.132	-0.022	-0.025	-0.033	-0.21	-0.52	
12	0.167	-0.089	0.015	-0.603	0.079	0.131	-0.061	-0.043	0.428	-0.07	0.135	0.065	0.079	-0.521	-0.283	
13	0.265	-0.064	-0.439	-0.75	-0.251	0.117	-0.223	-0.012	0.167	0.031	-0.01	0.109	0.214	-0.427	-0.362	
14	0.07	0.279	-0.132	-0.271	-0.188	0.131	0.061	-0.17	0.067	-0.063	-0.101	0.048	0.088	-0.092	-0.509	
15	0.445	-0.019	-0.378	-1.01	0.184	0.127	-0.308	-0.15	-0.021	-0.439	0.061	-0.145	0.014	-0.207	-0.636	
16	-0.118	-0.142	-0.08	-0.06	-0.252	-0.134	-0.204	-0.039	-0.003	-0.02	-0.043	-0.097	0.006	0.284	0.082	

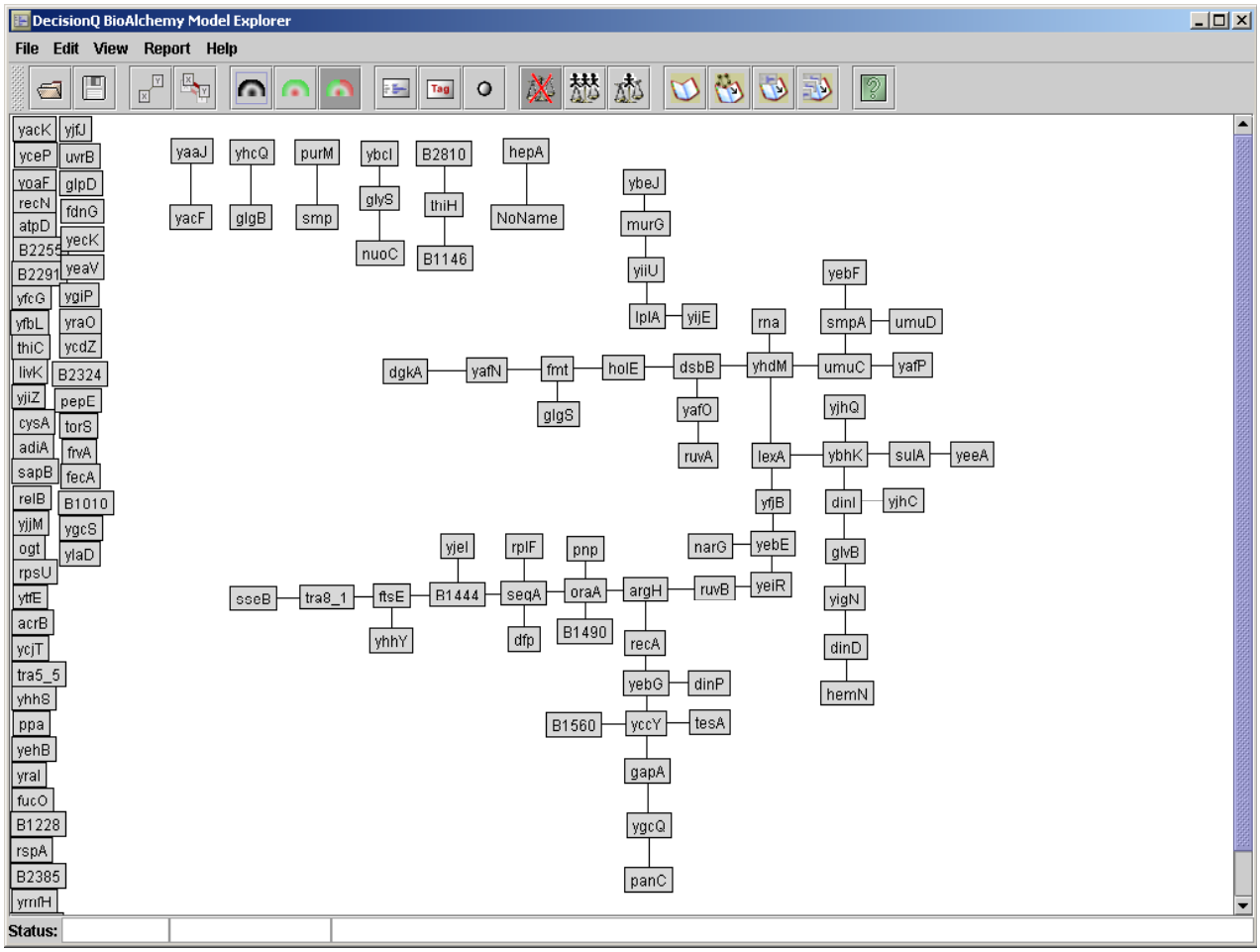
ecoli

Ready

Having selected the data, a fully automated process will continue until a full model is presented, or the user can stop each part of the process to manually change parameters in order to leverage particular domain expertise. The software begins by categorizing the data and 'binning' in accordance with the default settings; the data is then passed seamlessly to the Modeler for automated model development. Once the software has mapped all the complex

correlations and causality in the data a graphical model is presented in the Explorer (as illustrated in Figure 2). This whole process takes only minutes.

Figure 2: Base case model of the data presented in the "Explorer"



The display illustrates conditional dependence between variables and the pathways existing in the genome. Notice that approximately half of the genes (the unattached variables on the left margin) were dropped by the software as not being co-regulated with any other to a statistically significant degree. Also notice the small groups of two and three variables at the top left that form their own isolated pathways.

When interpreting the resulting gene expression network, we observe that many genes (shown as nodes) belonging to the SOS response appear clustered. To enable the user to quickly and easily absorb the relevance of the data visualization, FasterAnalytics includes a coloring feature that shows gene expression. This allows researchers to quickly view changes as they propagate through the network following a hypothesis.

In the example below, co-regulation with gene “recA” is shown. The thick border indicates that this gene (node) is the target selected, and its color green indicates that we are interested in analyzing how other genes (nodes) behave when the target is ‘underexpressed’. The coloring of the remaining nodes is red if the corresponding genes are ‘overexpressed’, and green if the corresponding genes are ‘underexpressed’ with tint of color being a relative measure of expression level.

Figure 3: Gene “recA” is set to ‘underexpressed’ and the co-regulation of other genes can be observed.

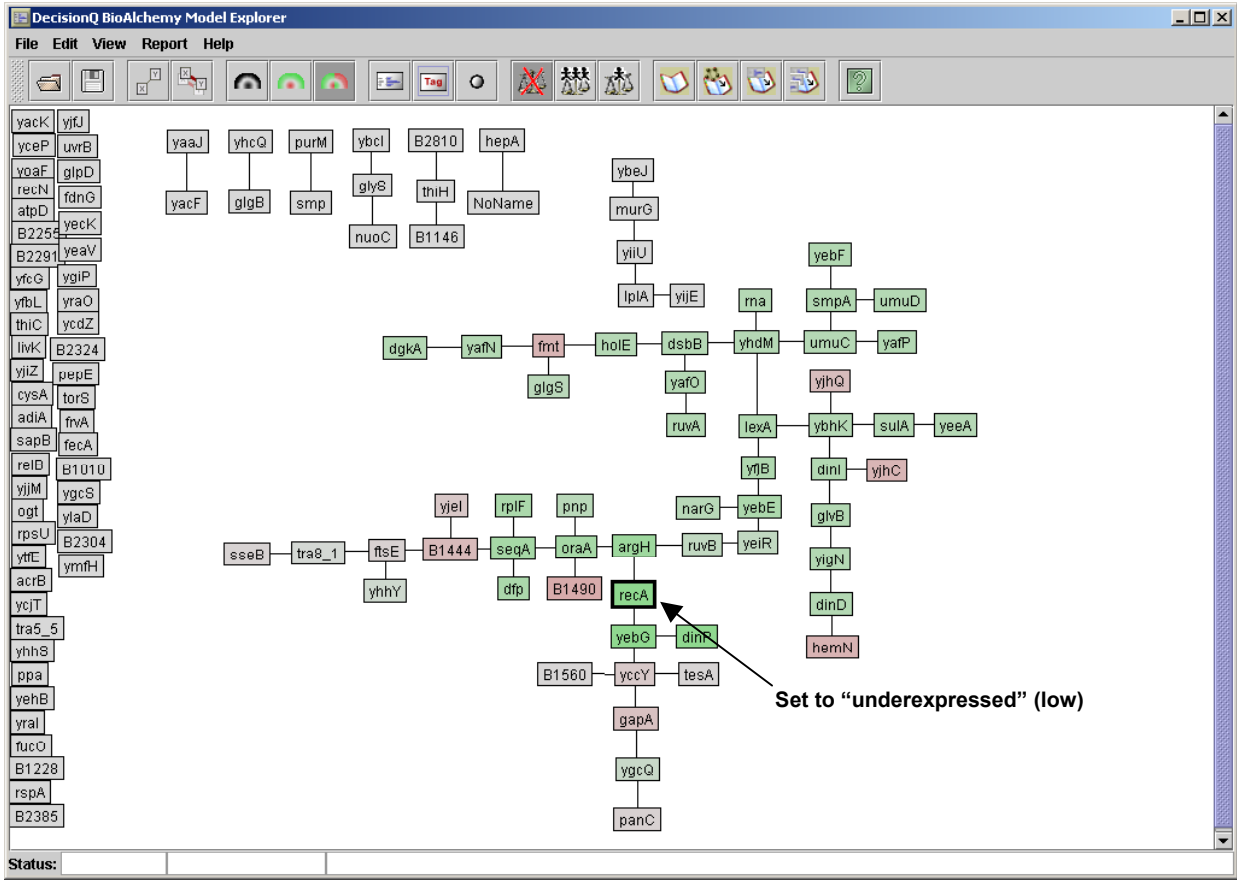
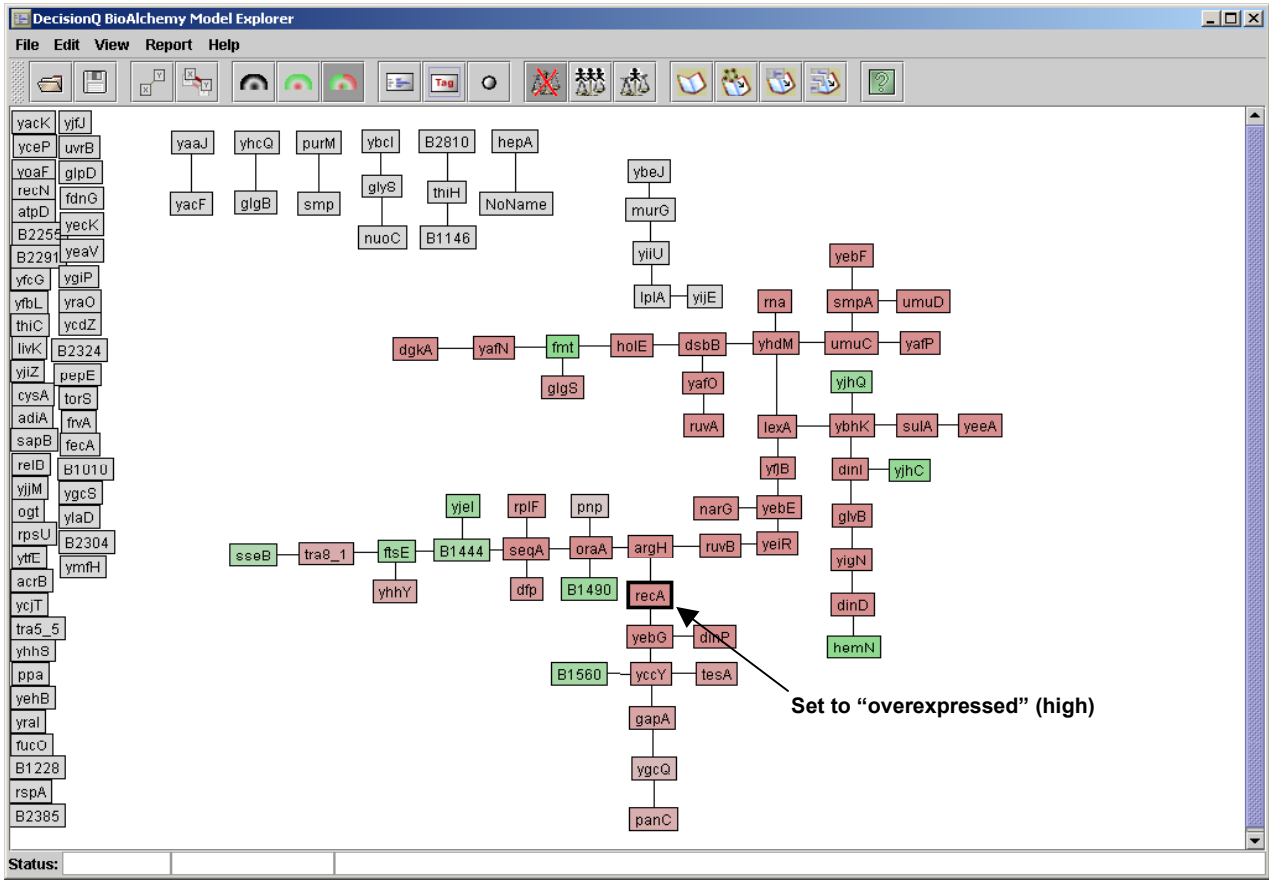


Figure 4: The complementary case analyzing the behavior of the different genes when gene “recA” is ‘overexpressed’.



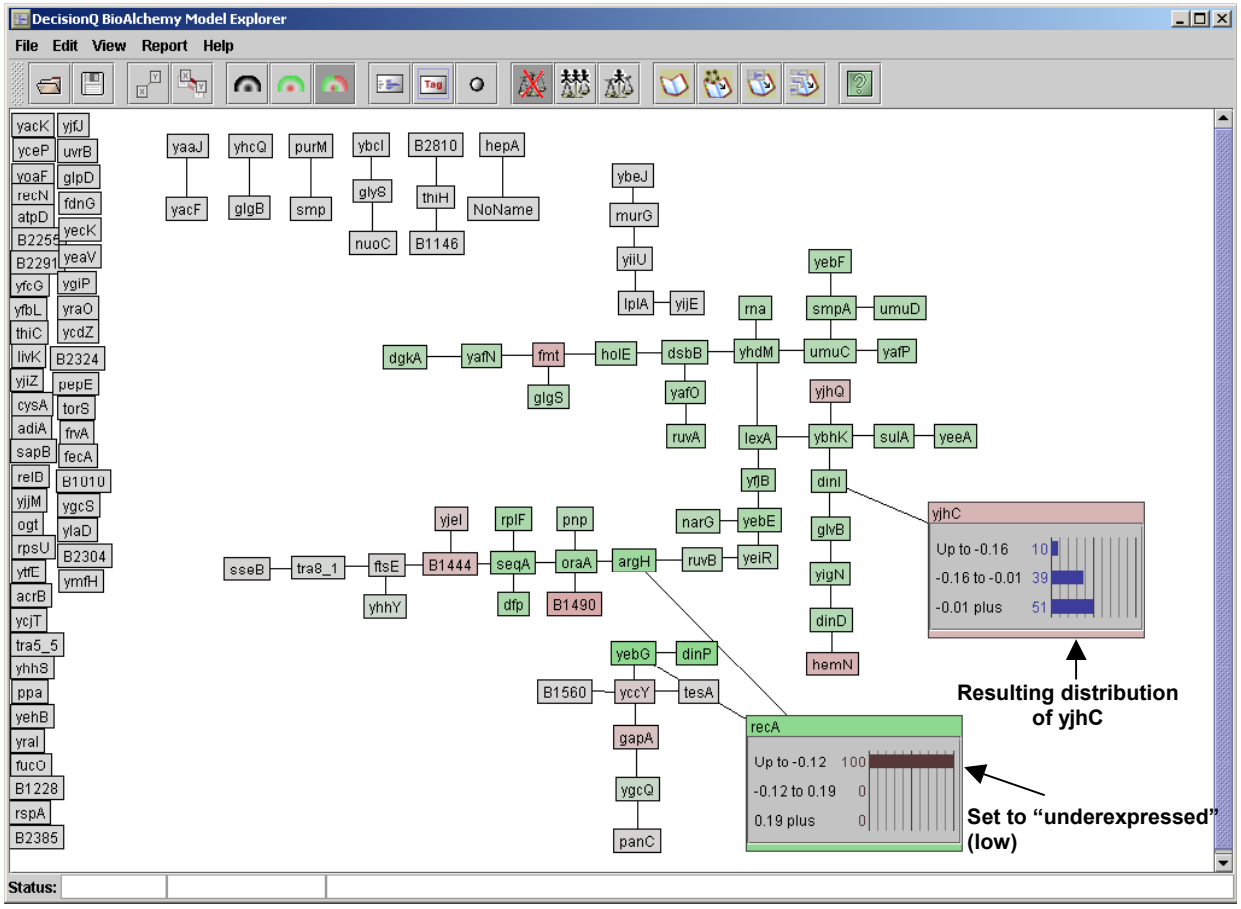
Compare the two models in Figure 3 and 4 above with the base level in Figure 2. It is also possible to select two or more gene targets simultaneously. The extent to which gene “recA” is co-regulated with other genes in its neighborhood is intuitive and clear. Critically, this can be used to narrow down the search for genes that are related to “recA”.

Each variable can be expanded using the 'View' menu or icons to show quantitative information about the relationships. The population data is displayed as "cases" with bars that represent the marginal probability distribution of each case. (Figure 5)

Suppose that we are interested in examining the relationship of "recA" with "yjhC". We first select these nodes and click "Graph" to display the states within these nodes. This can be done for as many nodes as we may choose. In the screen shot below the expression levels of "yjhC" are conditional on the 'overexpression' of "recA".

If we wish to test hypotheses, we can modify any node and see how our hypothesis affects the model. Notice how information flows through the network.

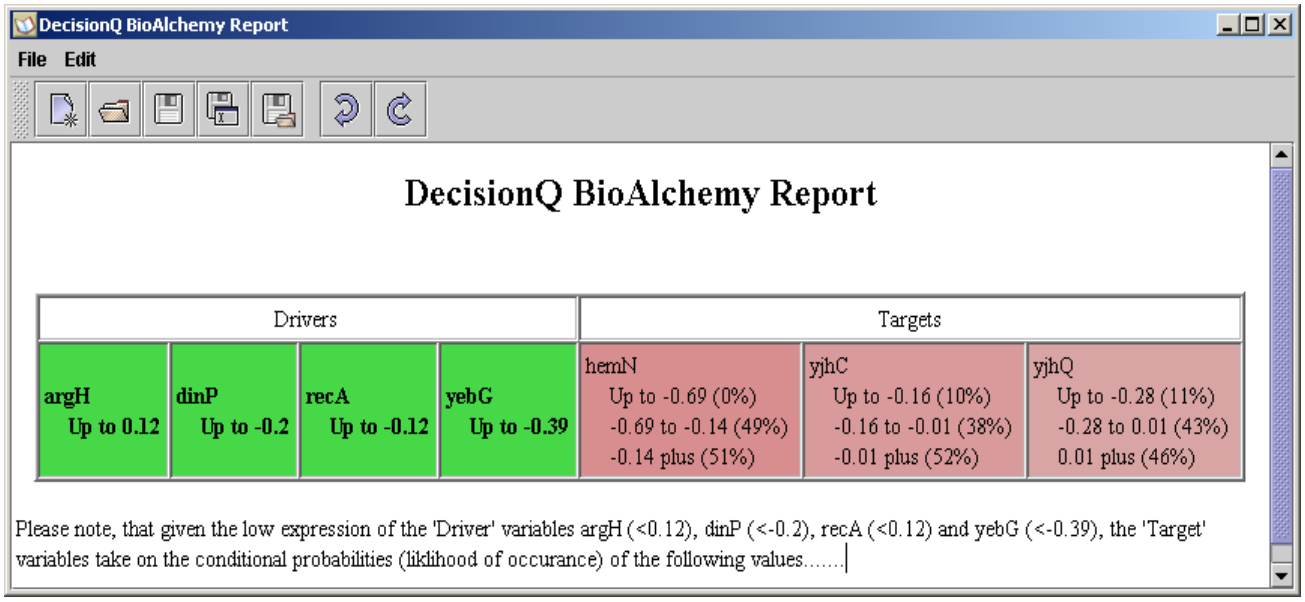
Figure 5: Quantitative information about "yjhC" when "recA" is 'underexpressed'.



The above data analysis is a basic example of microarray gene expression analysis. FasterAnalytics contains many potent features relevant to scientific knowledge discovery.

The Reporter module can be used to create a report that will show the conditional probabilities (or predicted likelihood) of any target variables, given the expression of any independent variable(s). Any part of the model visualization can be pasted into Reporter and then transferred into other applications. Figure 6 shows a sample report.

Figure 6: A sample report listing the probabilities of 3 random target variables given 4 driver variables.



DecisionQ sells predictive modeling software and complementary professional services. Alternatively, components from FasterAnalytics can be integrated into third party applications as part of broad data management and analysis platform.

If you have any further questions or would like to schedule a more detailed demonstration in person or over the web, please contact us.

DecisionQ Corporation
3726 Connecticut Ave NW
Suite 519
Washington, DC 20008
www.decisionq.com
Phone: 415-254-7996
Fax : 415-276-6356
Email: info@decisionq.com

1. Source of Ecoli data: J. Courcelle, A. Khodursky, B. Peter, P. O. Brown, and P. C. Hanawalt. Comparative gene expression profiles following uv exposure in wild-type and sos-deficient escherichia coli. Genetics, 158:41{64, 2001.